BAB II
TINJAUAN PUSTAKA

2.1 Pilkada Serentak

Pemilihan kepala daerah merupakan sebuah pesta demokrasi yang dilaksanakan setiap 5 tahun sekali. Berdasarkan berbagai sumber, pemilihan kepala daerah secara serentak ini sudah dilaksanakan mulai dari tahun 2015. Setiap penyelenggaraan pemilihan kepala daerah serentak ini banyak sekali timbul pro dan kontra. Pro dan kontra tersebut tidak hanya terjadi didunia nyata namun juga terjadi di dunia maya atau media sosial seperti *twitter*.

2.2 Twitter

Twitter telah menjadi sebuah media sosial yang banyak digunakan oleh para pengguna media sosial untuk berkomunikasi. Pengguna twitter dapat mem – posting *tweet* dengan batas hanya 140 karakter *tweet*. Twitter diluncurkan oleh Jack Dorssey padda tahun 2006. Berdasarkan portal web *techno.id* pengguna aktif media sosial twitter sebanyak 302 juta pengguna aktif yang 80 persennya berasal dari perangkat mobile dengan rentang usia penggunanya berusia 18 – 29 tahun sebanyak 37 persen dan 25 persen berusia 30 – 49 tahun[1].

Untuk dapat menggunakan platform media sosial twitter, para calon pengguna diwajibkan untuk registrasi atau mendaftarkan diri menggunakan *email* mereka. Setelah menginputkan data diri pribadi dan berhasil melakukan proses pendaftaran, pengguna bisa melakukan *login* dan menggunakan berbagai fitur yang terdapat di platform media sosial twitter. Seperti menuliskan status *(tweet)*, memberikan komentar terhadap status orang atau memberikan balasan pesan *(reply)*, melakukan sebuah *retweet* dari status setiap pengguna lain*(retweet)*, mengikuti pengguna lain *(follow)*, serta fitur lainnya. Dalam penelitian kali ini fitur *tweet* yang akan digunakan untuk mendapatkan sebuah data dari platform media sosial twitter. Dengan menggunakan API yang disediakan oleh twitter.

2.3 API Twitter

Twitter API merupakan sebuah layanan dari platform twitter yang memberikan kita izin sebagai developer untuk mengakses data atau informasi dari twitter. Untuk bisa menggunakan fitur tersebut *user* diharuskan untuk mendaftarkan diri menjadi *developer* terlebih dahulu pada situs http://dev.twitter.com menggunakan akun twitter yang telah dibuat sebelumnya. Setelah itu pengguna bisa *login* sebagai *developer* twitter. Selanjutnya pengguna bisa membuat *App* untuk mendapatkan *consumer key, consumer secret, access token,* dan *access key token* untuk digunakan dalam mendapatkan akses data dan informasi dari twitter.

Dalam memudahkan tugas *developer* mengakses data serta informasi dari twitter, tersedia banyak *library* yang digunakan. *Library* yang akan digunakan dalam penelitian ini adalah *twitterscraper* berbasis bahas pemrograman *python*. Untu kbisa menggunakan *library* tersebut diharuskan untuk menginstall *library* tersebut dengan menuliskan keyword *"pip install twitterscraper"* pada direktori *file* python tersebut.

2.4 Data Mining

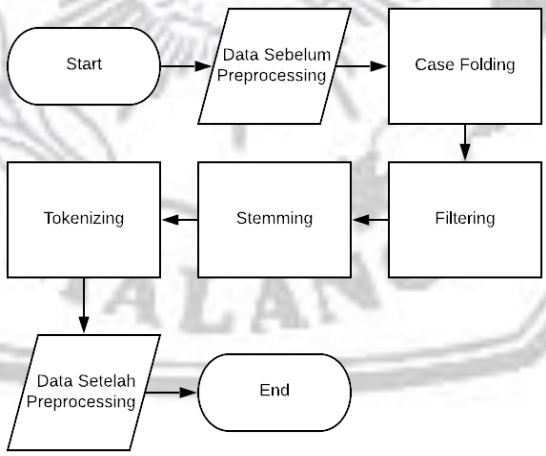
Data mining adalah sebuah proses untuk melakukan teknik pembelajaran komputer untuk melakukan analisis dan ekstrasi mengenai pengetahuan secara otomatis yang meliputi pengumpulan dan pemakaian data hostori dalam data yang berukuran besar. Adapun tujuan dari data mining ini sendiri ialah untuk mencarai *pattern* atau pola dari sebuah data dengan menggunakan metode dan algoritma yang menghasilkan suatu informasi atau pengetahuan yang bisa menjadi pertimbangan dalam pengambilan keputusan untuk ke depannya. Pengetahuan yang dihasilkan harus bersifat sesuatu yang baru, bermanfaat, dan mudah untuk dipahami.

2.5 Analisis Sentimen

Analisis sentimen adalah salah satu contoh dari penelitian teks mining. Analisis sentimen merupakan riset secara komputasional dari dua hal, yaitu opini dan emosi didalam sebuah teks. Adapun tujuan dari analisis sentiment ini ialah untuk mengetahui pendapat atau opini seorang pengguna media sosial terhadap suatu topik pembahasan. Respon yang dikeluarkan oleh pengguna terhadap suatu topik bahasan bisa dikelompokkan menjadi dua yaitu, positif dan negatif. Sehingga ketika melakukan sebuah penelitian berbasis analisis sentiment kita bisa mengetahui seperti apa respon pengguna terhadap topik bahasan tersebut.

Analisis sentimen dapat dilakukan dalam 3 tingkatan objek yang diteliti yaitu, pada tingkat sebuah dokumen, kalimat, dan aspek[11]. Penelitian ini dilakukan dalam tingkatan analisis sentiment dengan menggunakan kalimat sebagai objek untuk di teliti. Sumber data atau kalimat *(tweet)* tersebut berasal dari platform twitter yang telah melalui proses *crawling*. Hasil dari penelitian analisis sentiment dapat digunakan untuk pengambilan sebuah keputusan.

2.6 Preprocessing



Gambar 2.1 *Alur Preprocessing*

Proses *preprocessing* merupakan tahapan untuk memberikan data sehingga layak untuk diolah dan diproses pada tahap selanjutnya. Umumnya *preprocessing* dilakukan untuk mengeleminasi data yang tidak sesuai atau mengubah menjadi

bentuk yang lebih layak untuk di proses[12]. Dalam proses mengubah menjadi sebuah bentuk yang layak untuk di proses, dalam tahap ini terdapat *library* yang digunakan yaitu *library* sastrawi. Berikut beberapa tahapan pada proses *preprocessing* :

1. Case Folding

   Pada tahap ini terdapat sebuah proses untuk merubah huruf kapital menjadi huruf kecil. Cara kerja dari proses ini adalah memproses huruf alfabhet dari “a” hingga “z” saja. Sehingga karakter selain huruf tersebut akan dihapus[13].

2. Filtering

   Dalam tahap ini dilakukan proses pembuangan untuk sebuah kata yang tidak diperlukan dan mengambil kata yang di perlukan untuk digunakan dalam proses selanjutnya. Terdapat dua proses dalam tahap ini, yaitu sebagai berikut :

   a. Stopwords

      Stopwords merupakan kosakata yang bukan termasuk kata unik atau ciri pada suatu dokumen dan tidak menyampaikan pesan apapun secara signifikan pada teks atau kalimat[14]. Kosakata yang dimaksud seperti kata penghubung dan kata keterangan yang bukan merupakan kata unik, misalnya “dari”, “akan”, dan sebagainya.

   b. Punctuatioan removal

      Punctuational removal adalah proses untuk menghapuskan angka dan tanda baca yang terdapat didalam sebuah tweet[15].

3. Stemming

   Proses ini dilakukan untuk mendapatkan kata dasar dengan cara menghilangkan awalan, akhiran, sisipan, dan *confixes* (kombinasi awalan

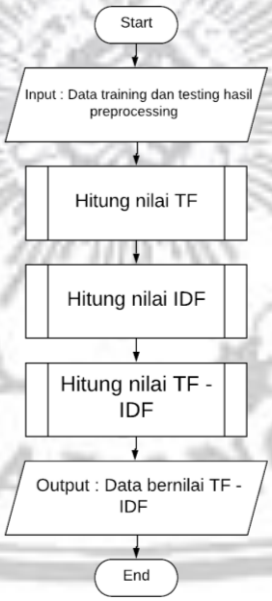
dan akhiran). Pada tahap ini akan menggunakan algoritma *nazief and adriani's stemmer*[16].

4. Tokenizing

Dalam tahap ini akan dilakukan proses untuk memilah isi teks sehingga menjadi satuan kata. Dalam proses ini prinsipnya untuk memisahkan setiap kata yang menyusun suatu dokumen.

2.7 Term Frequency – Inverse Document Frequency

Data yang telah melalui tahapan *preprocessing* siap untuk diolah. Pada data mentah tersebut akan dilakukan proses pembobotan pada setiap kata *(term)* dan memberikan hasil akhir berupa bobot TF – IDF. Hasil dari pembobotan ini yang akan digunakan dalam proses klasifikasi. Alur proses bisa dilihat pada gambar di bawah ini :



Gambar 2.2 *Alur proses TF – IDF*

Term Frequency – Inverse Document Frequency adalah statistik numerik yang menunjukan pentingnya kata pada dokumen. Umumnya TF – IDF digunakan

untuk menghitung bobot pada pengambilan informasi. TF – IDF dapat ditunjukkan dengan 3 cara [17], yaitu sebagai berikut :

1. Menggunakan nilai biner, yaitu diberi nilai 1 untuk kata yang terdapat pada dokumen dan nilai 0 terhadap kata yang tidak muncul pada dokumen. Pada konsep ini kata tidak dimasukan kedalam perhitungan.
2. Menggunakan nilai frekuensi kemunculan kata secara langsung untuk menjadi TF.
3. Menggunakan nilai pecahan *term* yang telah dilakukan normalisasi, dengan rumus sebagai berikut :

$$TF(t,d) = 0{,}5 + 0{,}5 \times \frac{f(t,d)}{\max\{f(w,d) : w \in d\}} \quad (1)$$

Berdasarkan rumus diatas dimana (t,d) merupakan frekuensi kata *t* muncul pada $\max\{f(w,d) : w \in d\}$. Sedangkan $\max\{f(w,d) : w \in d\}$merupakan frekuensi maksimum dari *term* lain pada dokumen *d.*

Sedangkan untuk menghitung nilai *IDF* digunakan rumus sebagai berikut :

$$IDF(t,d) = log\frac{N}{Df(t,d)} \quad (2)$$

Pada rumus diatas dimana *N* merupakan jumlah dokumen dan *Df (t,d)* sebagai banyak dokumen dalam kumpulan dokumen *D* yang mengandung *term t*. Namun bila *term* tidak muncul, maka terdapat nilai 0 pada pembagian, sehingga perlu penanganan untuk mengganti nya menjadi $1 + Df(t,d)$.

Untuk menghitung nilai dari *TF – IDF* menggunakan rumus sebagai berikut :

$$TF - IDF(t,d,D) = tf(t,d)x\ idf(t,D) \quad (3)$$

Berdasarkan rumus tersebut akan didapatkan nilai yang dijadikan sebagai pembobotan kata pada saat dilakukan pengelompokkan kata atau *term*.

2.8 Seleksi Fitur

Dalam metode untuk analisis sentimen berdasarkan pendekatan menggunakan *machine learning* terdapat ruang fitur yang begitu besar. Dengan ruang fitur tersebut, akan dating sebuah masalah. Metode seleksi fitur memiliki peranan penting dalam analisis sentimen, sama hal nya seperti dalam *text mining* lainnya. Penggunaan metode seleksi fitur membantu untuk memahami atribut yang relevan untuk kelas tertentu serta meningkatkan akurasi klasifikasi[18].

Terdapat dua jenis metode seleksi fitur dalam *machine learning*, yaitu *wrappers* dan *filters*. Menurut Jhon, Kohavi dan Pfleger[19].

2.8.1 Wrappers

Wrappers merupakan akurasi menggunakan klasifikasi beberapa algoritma sebagai fungsi evaluasinya. *Wrappers* harus menguji klasifikasi untuk setiap fitur bagian yang akan dievaluasi, biasanya lebih banyak ketika jumlah fitur tinggi.

2.8.2 Filters

Berbeda dengan *wrappers*, *filters* melakukan seleksi fitur yang menggunakan fitur yang dipilih. Dalam mengevaluasi fitur, *filters* menggunakan matrix evaluasi yang mengukur masing – masing *class*. Metode *filters* terdiri dari *information gains, term frequency, particle swarm optimization, chi – square, expected cross entropy, odds ratio, the weight of ecidence of text, mutual information,* dan *gini index.*

2.9 *Particle Swarm Optimization*

Particle Swarm Optimization sering digunakan dalam peneltian karena, metode ini dapat memecahkan masalah optimasi dan sebagai pemecah masalah seleksi fitur[20]. *Particle swarm optimization* tidak memiliki operator seperti *crossover* dan mutasi. Baris dalam matriks disebut partikel. Setiap partikel bergerak dengan kecepatan setiap pembaharuan kecepatan dan posisi berdasarkan lokal terbaik *(pbest)* dan global terbaik *(gbest)*.

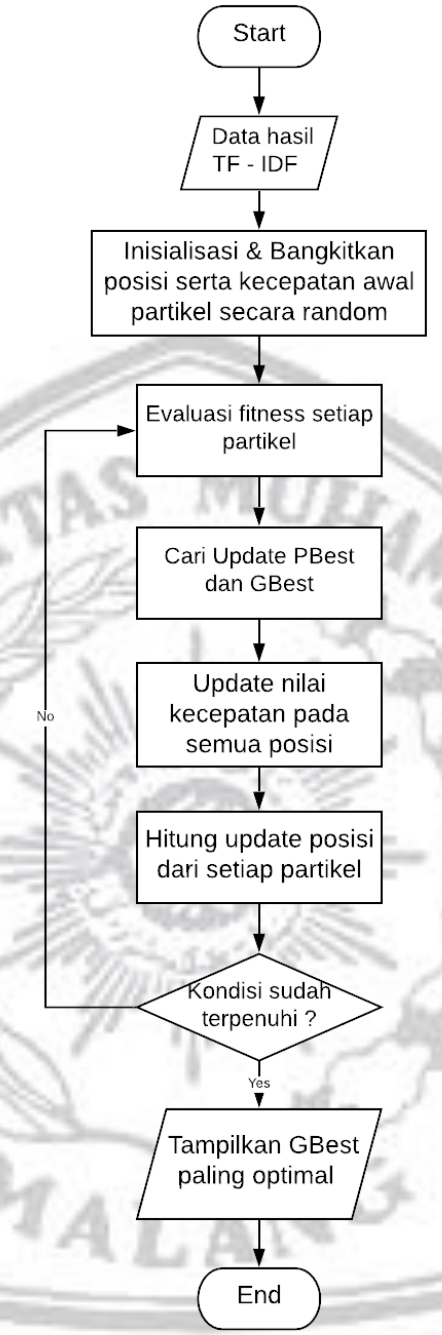
Particle swarm optimization merupakan sebuah konsep berbasis intelejen segerombolan yang dikemukankan pada tahun 1995 oleh Eberhart dan Kennedy[21]. Metode ini digunakan untuk meningkatkan akurasi terhadap atribut yang terdapat pada *naïve bayes classifier* dengan menggunakan persamaan sebagai berikut[22]:

$$V_i(t+1) = \omega V_{i(t)} + c_1 r_1 \left(P_{i(t)} - X_{i(t)}\right) + c_2 r_2 (P_g - X_{i(t)}) \quad (4)$$

Sekarang V_i merupakan sebuah kecepatan baru. Jadi, posisi pembaharuan partikel dengan kecepatan didefinisikan pada persamaan sebagai berikut :

$$X_{i(t+1)} = X_{i(t)} + V_{i(t+1)} \quad (5)$$

Berikut merupakan alur proses dalam melakukan perhitungan dari algoritma *particle swarm optimization*.



Gambar 2.3 *Alur Particle swarm optimization*

2.10 *Naïve Bayes Classifier*

Teorema *naïve bayes classifier* merupakan teorema yang mengacu pada konsep probabilitas bersyarat. Secara umum teorema *naïve bayes classifier* bisa dinotasikan pada persamaan berikut :

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (6)$$

Algoritma *naïve bayes classifier* merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk klasfikasi data uji pada kategori yang paling tepat. Metode ini merupakan salah satu metode untuk klasifikasi teks. Kelebihan metode ini yaitu, sederhana pada proses nya tetapi memiliki tingkat akurasi yang tinggi dibandingkan dengan algoritma klasifikasi yang lain. Terdapat dua tahap dalam klasifikasi *tweet*. Tahap pertama adalah melakukan pelatihan terhadap *tweet* yang telah diketahui kategorinya. Sedangkan untuk tahap kedua merupakan proses klasifikasi *tweet* yang belum diketahui kategorinya. Dalam algoritma *naïve bayes classifier* setiap dokumen di interpresentasikan dengan pasangan atribut "$a_1, a_2, \ldots\ldots, a_n$" dimana a_1 adalah kata pertama a_2 adalah kata kedua dan seterusnya. Sedangkan *V* adalah himpunan kategori *tweet*. Pada saat proses klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan *(Vmap)*. Adapun persamaan *Vmap* tersebut adalah sebagai berikut :

$$V_{map} = \underset{v_j=V}{\arg max} P(v_j) \prod_i P(a_i|V_j) \quad (7)$$

Nilai P(vj) dihitung pada saat data *training*, didapat dengan rumus sebagai berikut :

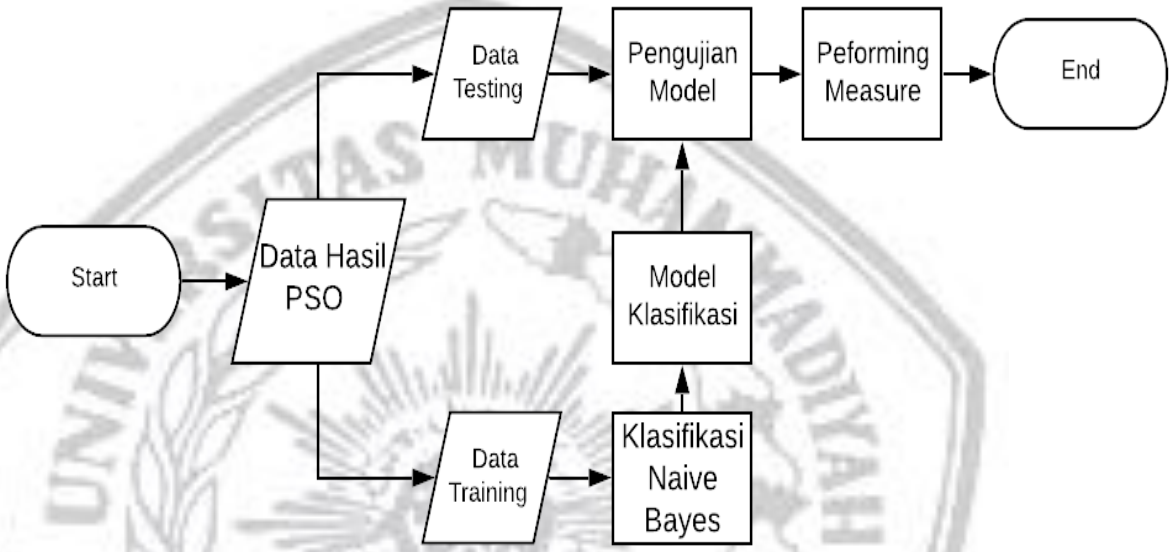
$$P(V_j) = \frac{|doc\ j|}{|\text{training}|} \quad (8)$$

Dimana |*doc j*| merupakan jumlah *tweet* yang memiliki kategori *j* dalam *training*. Sedangkan |*training*| merupakan jumlah *tweet* dalam contoh yang digunakan untuk *training*. Untuk setiap probabilitas kata a_i untuk setiap kategori $\prod_i P(a_i|V_j)$ dihitung pada saat data *training*.

$$P(a_i|v_j) = \frac{n_i+1}{|n+kosakata|} \quad (9)$$

Dimana n_i adalah jumlah kemunculan kata a_i dalam *tweet* yang berkategori v_j, sedangkan *n* adalah banyaknya seluruh kata dalam *tweet* dengan kategori v_j dan |kosakata| adalah banyaknya kata dalam data *training*.

Berikut merupakan alur dari proses yang akan dilakukan pada klasifikasi *naïve bayes*.



Gambar 2.4 *Alur proses klasifikasi naïve bayes*

2.11 Confussion Matrix

Confussion matrix merupakan sebuah cara dalam memberikan informasi dari perbandingan hasil klasifikasi yang dilakukan oleh sistem *(model)* dengan hasil klasifikasi sebenarnya. Terdapat beberapa istilah dalam *confussion matrix* yaitu. *True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)*. Berikut penjelasan mengenai istilah tersebut :

1. True Positive

   Memprediksi data positif yang diprediksi benar.
2. True Negative

   Memprediksi data negatif yang diprediksi benar.
3. False Positive

Memprediksi data negatif yang diprediksi sebagai data positif.

4. False Negative

   Memprediksi data positif yang diprediksi sebagai data negatif.

2.12 Sastrawi

Sastrawi merupakan sebuah *library* dalam bahasa pemrograman *python* yang dibuat oleh algoritma *nazief and adriani*. Algoritma tersebut merupakan aturan yang mengikuti pada aturan Bahasa Indonesia dan menjadi penentuan imbuhan yang diperbolehkan atau tidak. Imbuhan dikumpulkan menjadi imbuhan awal kata, tengah kata, akhir kata, dan penggabungan pada awal dan akhir kata[23].

Seiring berjalannya waktu *library* tersebut diperbaharui oleh algoritma *Confix Stripping*. Algoritma tersebut merupakan proses mengubah sebuah kata menjadi kata dasar dengan cara menambahkan kamus pada prosesnya. Kemudian *library* tersebut ditingkatkan menggunakan algoritma *Enchanced Confix Stripping*. Algoritma tersebut adalah pembaharuan dari algoritma sebelumnya yang memiliki fungsi untuk menyelesaikan kesalahn pada algoritma sebelumnya. *Library* sastrawi dikembangkan lagi untuk disempurnakan oleh *Modified ECS*. *Modified ECS* merupakan peningkatan dari algoritma *ECS*. Peningkatan tersebut dilakukan dengan menggunakan metode *corpus based* pada tabel penghapusan imbuhan[24].

2.13 Python

Python merupakan salah satu bahasa pemrograman tingkat tinggi yang memiliki sifat *interpreter, interactive, object – oriented,* dan bisa digunakan hampir pada semua *platform* atau *system* operasi seperti *linux, windows,* dan *mac*. *Python* cukup mudah untuk dipelajari karena *syntax* yang digunakan cukup mudah untuk dimengerti dengan banyak *library* yang tersedia untuk memudahkan pengoperasiannya, cocok untuk memproses data yang cuku besar dengan efisien[25].

2.14 *Jupyter Notebook*

Jupyter notebook adalah *tools opensource* yang bisa didapatkan secara gratis di internet yang digunakan sebagai *interface* atau antarmuka pengguna dengan bahasa pemrograman *python*. *Jupyter notebook* merupakan *editor* dalam bentuk *website* yang telah ter – *install* pada *localhost*. Berisikan lebih dari 300 *package python* untuk proses analisis data[26].